DOCUMENT RESUME

ED 249 290 TM 840 650

AUTHOR Welch, Wayne W., Ed.
TITLE Educational Evaluation: Recent Progress, Future Needs. Proceedings of the Minnesota Evaluation Conference (May 1980).
INSTITUTION Minnesota Research and Evaluation Center, Minneapolis.
PUB DATE Feb 81
NOTE 55p.
PUB TYPE Collected Works - Conference Proceedings (021) -- Information Analyses (070)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Educational Testing; *Evaluation; Evaluation Methods; *Evaluators; *Professional Occupations; Program Evaluation; Standards; Training; Trend Analysis
IDENTIFIERS Evaluation Problems

ABSTRACT

These proceedings include four papers on the status of evaluation Michael Scriven ("Evaluation: The State of the Science") examines the general picture of evaluation, identifies a number of trends in the field, and presents several new evaluation concepts. Included are discussions of "responsibility evaluation," "value phobia," the worth/merit distinction, apportioning, and the macro-micro evaluation distinction. Nick Smith ("The Progress of Evaluation: Rounding the First Bends in the River") argues that the field is insecure; it has a short history, is based on a small sample of phenomena, covers only a small part of a complex field, and is influenced by rare events. In spite of these concerns, he believes that evaluation is developing a new professional identity. W. James Popham ("Crumbling Conceptions of Educational Testing") admonishes evaluators to recognize the changes that have been occurring in the testing world. He presents five commonly-held but false conceptions which threaten testing and evaluation programs, and offers advice on how these threats can be addressed. Daniel Stufflebeam ("Standards, Research, and Training: Three Priorities for Professionalizing Educational Evaluation") asserts that, in order to build a more stable base of knowledge, work is needed in three areas: reseach, training, and professional standards. Standards are needed as ethics of conduct, research is required to understand the complex structures and relationships of the discipline, and training is vital to ensure the professional stature of the field. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. W. Welch

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Copyright© 1981 by Minnesota Research and Evaluation Center.

Educational Evaluation:
Recent Progress, Future Needs

Proceedings of the
Minnesota Evaluation Conference
May, 1980

Wayne W. Welch
Editor
University of Minnesota

Published by the Minnesota Research and Evaluation Center
February, 1981
Minneapolis, Minnesota

TABLE OF CONTENTS

		Page
Introduction and Overview	Wayne W. Welch	1
Evaluation: The State of the Science	Michael Scriven	6
The Progress of Evaluation: Rounding the First Bends in the River	Nick L. Smith	19
Crumbling Conceptions of Educational Testing	W. James Popham	30
Standards, Research, and Training: Three Priorities for Professionalizing Educational Evaluation	Daniel L. Stufflebeam	37
About the Contributors		50

Acknowledgements

The Minnesota Evaluation Conference and this report of its proceedings was greatly aided by the contribution of Ted Kellogg, Professor of Educational Foundations and Carmen DeGidio, Research Assistant. Professor Kellogg assisted in planning the conference and was responsible for pre-conference publicity. Ms. DeGidio served as copy editor and consultant during the development of this report.

Appreciation is also expressed to the College of Education and the Foundations of Education Department for their assistance with and support of the conference.

INTRODUCTION
by Wayne W. Welch

The profession of educational evaluation has experienced remarkable growth during the past 15 years. There are journals, annual meetings, training programs, graduate students, job opportunities, and national organizations whose major focus is the theory and practice of evaluation. In just a few short years, a discipline seems to be emerging where none existed before.

The factors responsible for this growth are varied, and the prospects for the future of educational evaluation are unknown. But the flurry of activity does suggest the need for careful scrutiny — evaluation, if you will. Where are we as a profession? How did we get here? Where are we headed?

It was this set of concerns that defined the focus for the first annual Minnesota Evaluation Conference, which was held in Minneapolis during May of 1980. Well-known evaluation professionals were asked to examine their discipline and to prepare papers setting forth what they found. Their findings and conclusions were presented to a group of 250 evaluators during the day-long conference entitled "Educational Evaluation — Recent Progress, Future Needs." The conference was organized by the Minnesota Research and Evaluation Center and supported by a grant from the College of Education at the University of Minnesota. The invited addresses were given by the following persons.

Michael Scriven
Director, Evaluation Institute
University of San Francisco

Nick L. Smith
Director, Research on Evaluation
Northwest Regional Laboratory

W. James Popham
Professor and Director
University of California at Los Angeles

Daniel L. Stufflebeam
Director, Evaluation Center
Western Michigan University

The four speakers were asked to address the general topic of the status of evaluation, paying particular attention to recent developments and to the future of evaluation as a discipline. The concept of the security of a discipline as presented by Boulding[1] was used as a further unifying theme and as a stimulus for additional perspectives.

[1]Boulding, Kenneth E. Science: our common heritage. *Science, 207*, 831-836.

Boulding defines a secure discipline as one whose knowledge is not likely to change much and an insecure discipline as one whose knowledge is less stable. Several characteristics of a field contribute to its relative security or insecurity. For example, insecure disciplines are characterized by the occurrence of rare or unfamiliar events, an inadequate history, complex structures and relationships, or a limited data base. These things usually are not part of a secure discipline. Thus, based on the above criteria, one could argue that our knowledge of classical physics, statistics, and literature is relatively secure while our knowledge of cosmology, paleontology, and human history is relatively insecure.

The speakers were asked to consider the security of evaluation as a discipline by identifying those things that they perceived to be stable and those things that they believed were likely to change in the near future. Nevertheless, this conceptualization was merely suggestive, and the participants were given considerable latitude in structuring their presentations. What follows are the perspectives on evaluation provided by three persons who have been a part of its growth for 15 years (Scriven, Popham, Stufflebeam) and one experienced but relative newcomer (Smith). Each of the papers approaches the task in a distinctive manner; together, they suggest a variety of insights about an emerging discipline and express cautious optimism about its future.

Perhaps, in some small way, this document will contribute to the security of evaluation by adding to the historical record of the movement.

Overview

Michael Scriven has brought his analytical mind to many complex problems in evaluation and has provided new insights, suggestions, and often, a new vocabulary. (Among his credits are formative, summative, goal-free, meta-evaluation, and modus operandi.) His article, "Evaluation: The State of the Science," continues that tradition as he examines the general picture of evaluation, identifies a number of trends in the field (good, bad, and neutral-but-notable), and presents several new evaluation concepts. Included here are discussions of "responsibility evaluation," "value phobia," the worth / merit distinction, apportioning, and the macro-micro evaluation distinction.

Scriven makes two other points in his paper that should generate considerable controversy among the theorists. First, he argues that the methodology of evaluation is secure but that the practice and political context are not. We have rigorous methods but, unfortunately, people don't usually use them. Second, he argues that evaluation, with its concern for multiple factors, should become a model for the social sciences. The present social science paradigm with its narrow focus on hypothesis testing, he says, is a failure in relation to evaluation.

As usual, Scriven is analytical, thought-provoking, and controversial.

Nick Smith's description of our evaluation journey down a metaphorical river (perhaps some would say "up the creek") is entitled "The Progress of

Evaluation: Rounding the First Bends in the River." He argues that the field is insecure according to Boulding's criteria: thus far, it has a short history, is based on a small sample of phenomena, covers only a small part of a complex field, and is influenced by rare events.

In spite of these concerns, Smith sees some hopeful signs. In his words, "Like shooting the rapids, living through such turmoil will result in greater security in the knowledge gained upon reaching the other side." Smith believes that evaluation is developing a new professional identity and offers some thoughts on what will, should, and can happen to help it successfully continue its journey.

He describes five vignettes, "scenarios of the possible," that offer glimpses of new methods to help the evaluation professional envision what can happen. They are descriptive, innovative, and suggest exciting possibilities for the future of the profession.

James Popham has been concerned with measurement problems for most of his professional career. In his paper, "Crumbling Conceptions of Educational Testing," he admonishes evaluators to recognize the changes that have been occurring in the testing world. Because much evaluation has been based on traditional testing procedures, it is important to understand the current climate in that area. In fact, he argues, evaluators must take an active role in order to maintain and improve the measurement opportunities in this country. Popham believes that tests do provide information that helps educational evaluators more effectively assess the merits of educational programs and therefore are essential to the profession. He presents five commonly-held but false conceptions which threaten testing and evaluation programs, and he offers advice on how these threats can be addressed.

While the other speakers in the conference tended to look in new directions for evaluation methodology, Popham effectively analyzed the changes in our traditional methods and called for action to improve what he perceives as a deteriorating situation.

In his paper, "Standards, Research, and Training: Three Priorities for Professionalizing Educational Evaluation," Daniel Stufflebeam admits that rated against Boulding's factors, evaluation is an insecure profession. This does not mean that it has no future, but rather that it is likely to be one filled with change.

In order to build a more stable base of knowledge, Stufflebeam believes that work is needed in three areas: research, training, and professional standards. Not only will such work ensure the continuation of the field, but it will contribute in important ways to its improvement.

Standards are needed as ethics of conduct, research is required to understand the complex structures and relationships of the discipline, and training is vital to ensure the professional stature of the field. Although Boulding ascribes no value judgment to secure versus insecure areas of human knowledge, it is clear that Stufflebeam believes that the security of evaluation is an important goal for the profession. He offers several well-reasoned arguments on how this desired goal might be achieved.

As a group, the four papers trace the recent progress of evaluation, admit that many changes are likely to occur, and offer a multitude of suggestions on what changes should take place to further develop and improve the profession and practice of educational evaluation. We hope that both the casual and the serious student of evaluation will find interesting and challenging images of knowledge to further their own development, as well.

EVALUATION: THE STATE OF THE SCIENCE

by Michael Scriven

I'm here to talk about two general areas. First, I will discuss the general picture in evaluation. Here I'll draw on some discussions that occurred at the AERA symposium, called "2001: An Evaluation Odyssey," where Eva Baker, James Popham, and Dan Stufflebeam were panel members. I will let Dan and James speak for themselves today, but I'd like to add one or two small points. I will also talk about some things Eva mentioned, which I think are very important. In the second part of the paper, I will discuss the logical foundations of evaluation and present some important issues that generally aren't in print.

To begin my discussion of the general picture of evaluation, I want to distinguish between the pre-science and the science of evaluation. The state of the pre-science — that is, the practice of relatively disciplined and sensible evaluation as illustrated in ***Consumer Reports*** and in a number of enthusiast magazines such as ***Car and Driver*** — is pretty good. It's rather highly developed in such areas as high fidelity equipment and automobiles; however, it's deteriorating slightly in others such as that in ***Consumer Reports***. The echo of this decline is the general thesis I want to stress to you: progress is no more inevitable in the field of evaluation than it is in civil rights. This is unlike, say, the field of mathematics.

When we are part of a new discipline, we always feel there is no way to go but up, and for a while, this is true. But in pre-scientific evaluation and in the science of evaluation, especially in the educational area, we have reached a point of potential stagnation. There are two reasons for this. First, there are people who have been in this area long enough to set it up as turf: the field has become "owned" by them. Thus, new movements are seen as challenging turf control, new methods are seen as intellectually threatening, and various other political problems begin to take important parts in the development of the subject. Second, in a new discipline, we get an "inversion of expertise" situation. The first people in the field were not trained in it, since such programs did not exist. Nevertheless, these persons are the senior staffers, and despite the fact that they often know less about the subject than the younger members, they tend to resist upgraded practices because of the implicit threat to their expertise.

The state of the science of evaluation is worrying now, and the state of the pre-science is often pretty primitive because the connections to the science are resisted by complacent pre-scientists. Accreditation is one example. The purchase of large and expensive hardware such as computers is another, more surprising, one. This is not usually done on sensible and solid kinds of evaluative foundations, as recent studies have shown. The big companies purchasing the million-dollar-plus computer installation mostly do it on a wing and a prayer. Someone is told to do it, and they do it on the basis of having good relations with one of the vendors, on having read somewhere that a particular machine "looks good," or on seeing in the enthusiast magazines that this one is the "in" thing. This is not a very

encouraging thought, especially when you realize that most of the federal agencies that you support operate in just that way.

The state of evaluation as a science is a little schizophrenic. It is secure with respect to methodology, which is a crucial point. But with regard to the practice and the political context, the situation is quite variable. Some other fields also have this problem. Medical research, for example, is an area where the methodology is very secure, but the practice is not good. Someone from the Office of Technology Assessment approached me after the AERA session on product evaluation and said they had studied the objective investigations of various medical practices. One of their discoveries was that the ***Journal of Neurosurgery,*** which is a leading journal in that field, had published 5800 articles since 1977, one of which was a controlled study. Next time someone has a knife in your brain, try not to think of that. Reading research is another area; too rarely do we find control group studies where they would be feasible and decisive. Nevertheless, there are areas where it is perfectly straightforward, though not easy, to use rigorous methods. Unfortunately, people usually don't use them. We see that big gap between the practice and the available methodology that needs to be bridged more than it has been.

Now I will discuss some trends — good ones, bad ones, and what I call "neutral-but-notable" ones — to support the general theme concerning the state of the science. First, there are several things we **shouldn't** be doing, but may be doing more often. One recent tendency Eva Baker noted was the substitution of what might be called "implementation checks" for evaluation. Running an evaluation shop with a large number of contracts, she has a good finger on this pulse. It is clear that we are seeing more legislation and directives (at the state department level, as well as at the federal level) in which the evaluation mission is only to find out whether the money was used in legitimate ways, with or without success. This "fiscal audit" approach is very attractive to persons who want to make themselves look good, since it's easy to show that an agency or a contractor didn't steal the money. Implementation, in this loose sense, is something one can easily control. Results cannot be as easily controlled, except through efficient design and successful practice or theory (which we lack in many areas of evaluation). The problem is that children in special education don't benefit from knowing that the money allocated to them was not stolen; they benefit from learning how to read. Implementation checks substitute a process indicator for an outcome indicator; and in general, it's an unfortunate practice.

Another concern, which was raised by Dan Stufflebeam at the symposium, is the trend toward self-evaluation. Such research replaces external evaluations for cost and other reasons. But if we are not careful, such a move could represent a deterioration of objectivity and could threaten the scientific status of evaluation. The obvious reasons are the potential subjectivity, invalidity, and lack of credibility of self-evaluation. There are good features of self-evaluation, and it should be a component of both formative and summative assessments. Nevertheless, it is insufficient by itself.

Finally, under the "bad" heading, is the influence of the dead hand of the old social science model on evaluation — a refinement of the inversion of expertise problem. We now have many persons who are officially called evaluators (particularly in education, but also in the health care, the criminal justice, and the urban renewal fields), who are social science trained. Because they write the RFP's to which jobbers respond and they monitor the contractors' performances, they greatly constrain what is done in evaluation. We are suffering now — and will suffer to a greater extent in the future — from the fact that with respect to evaluation, the social sciences were crippled at birth by the doctrine of value-free social science.

The social science doctrine makes the notion of scientific evaluation a contradiction in terms. Social scientists have always had difficulty swallowing the idea of doing evaluation scientifically; they were brought up to believe that value judgments had nothing to do with science and that such judgments were purely subjective and arbitrary. And yet here they are doing social science based evaluation. There are various ways to wiggle around on this pin, but it's not a comfortable spot to be in, and wiggling doesn't do much good. The situation is serious: a basic inconsistency exists between the assumptions of the approach and the practice of evaluation, and most evaluation work is guided by persons who were trained to regard it as taboo. We must become more self-conscious about the development of evaluation as a new discipline. If you like, it's a social science, but in a new sense of it. In fact, as I suggest later, evaluation is best thought of as a new model for other social sciences, rather than the other way around.

Now let's look at some notable changes that will affect evaluation, especially in education, but which are not intrinsically good or bad. Eva Baker observed that evaluators' value seems to be proportional to their extra-evaluation knowledge. (I've noticed the same thing in Philosophy.) We have to look very closely at the type of training evaluators get, and we should avoid purely methodological or quantitative instruction. An approach I would like to see more commonly used is the hybrid team, which consists of evaluation methodologists and subject matter experts. There might be one, two, or three persons representing each expertise on the team. Nevertheless, local knowledge with methodological expertise is a winning combination.

Another change concerns the move toward legal involvement in evaluation. The courts have moved in, and they've moved in heavily. In Florida, for example, the courts went in to review and critique the testing programs. They intervened, of course, because people have been asking them to. Since it's not clear what their track record will be, I've included it in the neutral-but-notable category.

In contrast to the area of testing, we have not seen long parades of expert witnesses in evaluation; but we may still see it. In Minnesota, there resides one of the two people in the country who are supremely competent as expert witnesses on the reliability of lie detectors. I spent some time talking to him about the role of expert witnesses. That's a fascinating question; of course, the bottom line is that they are shockingly unreliable. As the courts

become more involved in evaluation, we may find more activity by evaluators as expert witnesses. We need to ask ourselves if we are prepared for that. We should also examine the possibility of augmenting our educational programs with some legal training.

Another neutral trend is the way certain federal offices may become the leading forces in evaluation. It won't be USOE or the Department of Education. Nor will it be NIE, except for perhaps one or two projects like Nick Smith's Research on Evaluation project, which is a very small and decreasing part of the scene. They appear to have dropped the ball because evaluation is a hot political item. But some other things have come through. A few days ago, the Government Accounting Office (GAO) created a state of the art Institute for Program Evaluation. For a number of years, GAO has been looking better and better. They are moving into evaluation in a far more sophisticated way than USOE ever has. GAO is becoming like NIE recently has become in its own practice (as contrasted with the kinds of things it partially subsidizes). But GAO and perhaps the Congressional Budget Office and the Office of Technology Assessment are becoming good program evaluators. Maybe this entry should be in the hopeful category, but it's certainly notable.

There are several other neutral-but-notable things I think are important for the 1980's. They have not yet emerged, but when they do they will make quite a difference to educational evaluation. The first I'll call the death of the basic curriculum, as a result of pocket computers. As far as I can see, instruction in about half of the basics will be eliminated by the mid-80's; and instruction in all of them probably will be gone by the end of the decade — that is, if we act with the children's interests in mind.

This is something we should have seen earlier. It's always a betting game to predict how a certain technology will develop, but the direction in this area is fairly clear. Let me walk you through the scenario.

The introduction of the hand-held calculator has eliminated the necessity for learning complex computational skills, and in many cases, it has led to large changes in teaching. (That this is not universal isn't the issue; that's an issue of responsiveness.) It's clear that we can eliminate the teaching of many complicated computational skills. We may need to teach some basics so students can multiply single and double digits together, but we certainly don't have to go beyond that. In general, we should critically examine the arithmetic curriculum and get a little tougher about it. If we persist in making the development of paper computational skills a major part of it, we must realize that some students can't cut it. And the difficulty such persons would have enjoying or even succeeding in school would be unnecessary, since they could easily master the hand calculator. Ignoring this fact will cause problems, because parents won't sit still for it.

The hand-held calculator is the thin end of the wedge. We already have translation calculators into which one taps an English word or phrase, and the Spanish, French, or German equivalent comes out on the screen. Also, voice output translators recently have been introduced on the market. These machines speak out the Spanish or French equivalent of whatever

you enter in English. Of course, some basics are required, since you must be able to enter it.

Now let's consider how to eliminate reading and writing. It'll go like this. The first effective omnifont optical character reading program, developed by Kurzweil as part of a special program for blind readers, was bought by Xerox a few weeks ago. Kurzweil was doing fine, but now that Xerox will put big money into them, they will begin to pick up tempo. Thus, we have the optical scanner, into which one can feed any printed material. This material is then digitized and converted into a universal language that is then printed in Braille, typed, or spoken out. We need only to reduce them to hand-held size and lower their cost to produce a big change in education. This will take only about one-and-a-half more generations of memory technology, one of which we have already parented. So we can expect to be able to buy a little "super calculator" that has a reading cell on the back. As the cell is moved across a line, it speaks out what it is reading. Thus, we have eliminated reading in favor of speaking, and children already speak before they get to school.

Learning how to write, on the other hand, is partly eliminated by substituting typing for handwriting, which is done on the keyboard of the widget. The rest is eliminated by voice input. Voice input is a little tougher than voice output, but it's developing fast. A leading Japanese firm has demonstrated an experimental voice input word processor with a vocabulary of 100,000 words in Japanese. It s not hand-held size, but it's removed from it only by memory condensation and large-scale integrated chip design.

In sum, we are approaching a situation in which we will have our hand-held calculator at a cost of $150 to $200 (corrected to 1980 dollars) by the end of the decade. It will work in the following way. You speak into the machine, and it will do various things with the words. For example, it will type them out onto a tiny printout; for people who can't read, it will feed it to their widgets; it will speak out in another language; or it will call up a data base with the kind of information you are requesting. It can also be used to convert in the other direction by reading documents in old-fashioned print and speaking them out. I think that takes care of everything: arithmetic, reading, writing, researching, and foreign languages. It will all be there.

What will be the response of the educational establishment? Some of that is predictable; some is unclear. Teachers, for example, will probably become very nervous about what to do with their time. They also will rationalize a negative posture by saying that if we take this sort of stuff too seriously, children's brains will rot through inactivity. This is an incorrect response, since one can simply move further up the curriculum scale, with all of the appropriate Piagetian corrections.

The inevitability of success, despite opposition, will probably be due to the presence of the Bureau of Education of the Handicapped (BEH). The money in special education will get that widget made. Special education children cannot be forced to handle decoding problems they don't need in

order to get into and succeed in the mainstream. And the parents of regular kids will be screaming loudly if their kids are forced to spend six years mastering basics that special education kids are not required to learn. You can expect that to be the end of it. Of course, there will be a college track group where quaint old-fashioned skills such as reading and writing are taught for a while, but it's not clear how long we'll even continue that.

Some interesting issues are raised by this projection. For example, there is the problem of reading speed: you can read much faster than you can hear. This is an important point for continuing to train scholars to master the old skills, just as for many years they learned Latin. There are other issues of this kind; for example, the difficulty of analyzing. But this type of skill requires a ten-year time-out training period, and that's a very high price.

I don't want you to think that I'm blue-skying this; I'm deliberately discussing things that have been invented. We have, in pilot form, essentially all we need. And in most cases, we have large-scale, fully operating machines. Miniaturization plus some software is all we need to get us down to the widget. Voice input, in particular, still needs some software; but that's relatively simple, as we see from the Toshiba performance.

What are the implications for our profession? First, I think it's going to change our emphasis in educational evaluation. In evaluating basics programs two to five years from now, the key questions should be the "critical competitor" question; that is, why they have the program at all. Good evaluation is almost always comparative evaluation, and the comparison must always be made with the most viable alternative approach. Some clients don't like that. They would prefer you to ask: "Are you succeeding in teaching these kids how to read?" The answer to that question may be "yes," but the answer to the real evaluation question might be that they should not be teaching them that. Will you, as a professional, be willing to take the tough line and question the whole nature of the effort? Will you know enough to be able to say when the alternative is viable? Will you let your ignorance of the technology lull you into accepting the soft question (success) as your agenda instead of the hard one (justification)?

A second implication of these technological developments is that there will be radical changes in educational research. This will affect evaluation as one species of educational research. We may see the end of hard copy, we may no longer subscribe to journals, and we normally may not buy books. We almost can't afford books now, and there is no question that we won't be able to afford journals, since our salary will go down and the price of journals will go up. Instead, we will have a terminal that is connected to an on-line data base. Now, for about $600 or $700 you can get a TI Silent 700 computer terminal (which is also an electric typewriter), that hooks up through the telephone to one of the data bases. This will print out the abstracts, the tables of contents, the first paragraphs, or ultimately, the contents of every article that meets your interests from 6000 journals. In

education alone there are 6000 journals; if you have additional interests in the social sciences, this number is greatly increased. Since you will not subscribe to, let alone read, 6000 journals, and the computer will do an almost instantaneous search, it's a considerable improvement.

The cost of this computer service is zero in subscription terms. You pay for it to the extent that you use it, but you pay no entry or monthly fee. And you can get the on-line connect costs down to a very low level with a joint subscription through an institutional library.

This development will also affect our professional communication. For example, it will lead to the creation of what I call "time-spread conferencing." That is, people will not **come** to conferences; rather, they will be **tied** to them. The papers will be "read" or presented through communicating phone lines to word processors and computer terminals. Participants will have, say, forty-eight hours to read and react through their system. The responses will appear on the screens of other people's terminals, and then they will react, and so on. This "continuous publication interaction" is like correspondence chess; it eliminates travel costs (which is just as well, since people won't be able to afford travel anymore) and greatly reduces the pressures of immediate, face-to-face responding. You can take a few hours to carefully think up your reactions, and compose, edit, and revise them before you send them out across the phone line.

A third major occurrence I want to underscore is the changing role of public education. Mike Atkin, ex-Dean at Illinois and now Dean at Stanford, who perhaps knows as much about it as anybody, thinks that the end has come. He believes that there isn't any way public education will be able to survive the intervention of the Feds. It's perfectly clear in the Constitution that the Feds are not supposed to get involved in education; but it's also clear that through civil rights they have found, or have been forced into, a way around that. They now stipulate that special education teachers do certain things, such as call them in to consult and advise, meet with parents every so often, and so forth. They are in every classroom, and with the Title 20 requirements, they are in the personnel office. What this means is that public education is under the fire of every political pressure group, and we know some of the effects of that. It won't take a voucher scheme, only an income tax deduction to tilt the balance. We must face the possibility that it will become more serious. This may mean that the relative importance of private and public education will change radically. In Australia, for example, the mix is very different with perhaps 40% of the students attending private schools, yet it appears to be highly workable. We need to keep an open mind about how children are best served.

Now I'll turn to the "good" or should category of trends. Eva Baker has seen some signs that information technology may increase the amount of consumer-oriented evaluation, and that as a result the Standards project may have a greater effect. I believe that it's not likely to be a significant influence. Recall what happened with computer assisted instruction (CAI). It was in this same high technology area that possibly the worst

evaluation in the history of education occurred; in other words, a failure to seriously evaluate until much later in the Educational Testing Service (ETS) contract. One reason this happened was that people pulled scientific rank on the evaluators: the hot-shot engineers kept staring down the evaluators (who were lowly education or social science types) and got away with twenty-five million dollars before anyone would listen to what the evaluators were saying. This is not an ideal model for us, but it probably won't be any better in other high technology areas.

There are other factors, however, favoring an increase in consumer-oriented evaluation. One is the Nader influence, an orientation which the manufacturers don't buy. It's also obvious that Congress can be subverted from it, as we see in their attitude toward the federal agencies that are in charge of such activity. For example, their refusal to rebudget the agencies and their requirement that everyone be fired and then rehired. This reflects the widely-held views that consumerism (or environmentalism) has "gotten out of hand" and that it should be "gotten under control." In some respects, they're right. The lawsuit situation, for example, is at the point where the suits make no sense. Nevertheless, there is a great deal of pressure to kill off serious evaluation, and I'm not sure we'll survive it. We may only see the elimination of federal support for any kind of serious evaluation. Or perhaps GAO or the Congressional Budget Office (CBO) will save us; but perhaps they won't survive in their new form. There is a tremendous opportunity for doing something with the evaluation of educational material. I hope we can seize it, one way or another. Learner verification has been introduced and has reached legislative status in Florida and in a number of other states. I think there is some hope there, but it's not clear which way it will go.

There are several things I hope to see. One is that evaluation becomes a standard part of the curriculum in the 1980's. Such training and perspective must be made into a key component of every education degree and of every sub-area of specialization. This is a position that it must, on its merits, be accorded. It's senseless to train professionals in curriculum, teaching, administration, or as specialists in particular areas without evaluation; and yet this still happens. Of course, there is an entrenched opposition; that is, you must displace other courses to get evaluation classes in there. But the problem is that we are churning out educational administrators who don't know a damned thing about evaluation of teachers, supplies, or administrators; except, that is, for what they picked up by their noble attendance at conferences like this. But we shouldn't be in a position of forcing them to devote in-service time to something new graduates can perfectly well get as a part of pre-service training.

I hope to see a change in this area. We are seeing changes in a few departments, here and there, but it's a speck in the ocean. We can't continue to turn out school administrators who think that the way to evaluate teachers is to drop in on the classroom now and then and write something on a checklist. Research has shown that nothing you observe there — even if you watched every session — could possibly bear on the

question of the teacher's merit. The whole thing is ludicrous. And the really bad news is that administrators don't even know the argument for saying it's ludicrous. The situation is the same with reading specialists, who don't know how to evaluate reading material, and so on. Of course, they all do something that they call evaluating, but it's invalid and expensive.

I also hope to see us get all of the little goodies that make up evaluation into each specialty area; for example, cost analysis, ethical analysis, and assumption analysis. In cost analysis, we have to get beyond the awareness that cost is a problem and learn how to do it. Cost analysis is not a trivial exercise; it takes very hard work. The Smith project has gotten Hank Levin to write some materials on this. Since there are just a few small things around now, I list it as a hopeful sign.

I gave a paper in Greensboro at the Southeast Measurement Conference about a month ago in which I set out a fairly complete system for evaluating teachers. This system considers only teachers' ethical obligations and nothing else. Standard procedures for evaluating teachers usually do not consider these factors. But the ethical component, in both personnel and program evaluation, is enormous. We haven't really faced this. One reason is that we are very nervous about the idea of doing ethical analysis. We have to get un-nervous about it, unless we want to get rid of ethics. As long as ethics exist and are relevant to the treatment of human beings, we have to face them every time we do a program evaluation. The same kind of point should be made for assumption analysis. In both curriculum and program evaluation, we are constantly faced with a situation where identifying the pre-suppositions of the activities is the main key to a serious assessment. It's a quite a trick to do it, but we must.

I've just given you some ideas about what evaluation training programs should include. They also should include self-evaluation; that is, an evaluation of the training program itself. In fact, evaluation of training programs should become a paradigm of how to evaluate educational efforts.

The federal government, however, has essentially eliminated funds for serious training in evaluation. While they are employing more people in this area than ever before, I'm sure it's because of the political pressure and the idea that the turf is already controlled. For example, such programs might suggest that there are jobs when, in fact, the civil service wants to shuffle around people who are already in it to fill up slots that turn up. That's not a good approach.

I've talked about a number of things I believe are central to the conceptual foundations of evaluation, things that usually aren't mentioned. A couple of items I want to mention now are things I've written fairly extensively about. One is checklist methodology, which I think is the key methodology for evaluation. Checklist methodology may be contrasted with the theory-based (the application of theory) methodology of the social sciences and with what the term "applied social science" suggests. The idea that evaluation is applied social science means that there are some general-level social science principles that you can apply to a particular

evaluation problem. I don't think there are many; there are just investigative methods.

People have said that the Keller Plan is an example of applying Skinnerian learning theory to education. This idea turns everyone with humanities training off. But if you actually look at Keller Plan teaching, what are you looking at? You're looking at a form of teaching on which my college at Oxford was based when it was founded in 1453; namely, the tutorial/individual study approach. This was the original classic paradigm of humanities instruction. It didn't have much to do with Skinner at that point; the "application of behavioral research" nonsense was just nonsense. It was the application of a bit of common sense and practical wisdom that lately has acquired a new jargon. We need a different way of conceptualizing what we are doing in evaluation. Checklist methodology, not theory methodology, seems to be the kind of approach that is appropriate.

The apportioning issue, which I think Wayne mentioned, is also very important. We have to understand the differences among apportioning, grading, ranking, and scoring. These are the four predicates in evaluation, and only some of them are appropriate in a given case. It's like formative vs. summative evaluation; if you don't understand that difference (whether or not you use those terms), you are likely to start doing the wrong thing in a particular job. Apportioning — as opposed to grading, ranking, and scoring — has an important role that is closely connected to the decision-maker's need to allocate funds across programs, in contrast to his or her need to conclude that one is better than another.

The worth/merit distinction, I think, is tremendously important. Worth is defined as the value of an individual to the institution. This is in contrast to an individual who has merit, but no value to the institution because he or she draws no students. In other words, a teacher may be someone who in other respects would be valuable to the institution, but is not professionally competent, and hence not acceptable. To this I want to add the concept of "market value" so as to cope with the problem of differential salary scales for medical school faculty, the hard sciences, business, and the humanities. Differential "costs" (pay) to get comparably meritorious faculty, like the differences in worth to the system from differential enrollments ("internal worth"), are not a sign of differences in **true** worth. Rather, they are signs of differences in "external worth," and hence may call for different responses by the institution in order to get its mission accomplished.

Now I want to mention some things I've been working on lately. One involves the notion of responsibility evaluation. I began developing something along this line a number of years ago. It is called the School Evaluation Profile (SEP) for the San Francisco Schools Commission, and it involves a method for evaluating a single school site. It should tell you how good a particular school is, and by generalization, how good a district is. The SEP operates on the important pre-supposition that there is no point in evaluating schools with respect to variables they can't control. For example, we don't want to get into the business of determining schools' average socio-economic statuses or average IQ's. Such indicators of merit

suggest that if you want to improve the school, you should move it somewhere else. In responsibility evaluation, variables are counted only if they are controlled by one of the school's communities; for example, the parent community, or the district administrator group. Thus, it is a method of evaluation scoring where you consider all the variables that affect the school's performance given its fixed condition; for instance, its fixed intake of students, the fact that its budget is fixed within a certain range, and so on. Evaluation of the school in terms of how well it's doing given such constraints is responsibility evaluation. This is something we need to do more seriously. And it ties in, of course, with the ethical dimension of evaluation.

Another topic I've been working on lately is something I call "value-phobia." The early work on bias control led me to see more of the complex climate of circumstances in which evaluation tries to operate. We are all aware of evaluation anxiety — the threat produced by evaluation — but we haven't seriously looked at it in detail. We need to be more conscious of its origins. Sometimes it's a generalized threat causing insecurity about job performance. Sometimes it's based on the value-free social science doctrine; in other words, the beliefs that we shouldn't be in this business at all, that it can't be done objectively, and that it's just personal prejudice. There are other sources that we need to identify.

Valuephobia creates an enormous reactivity problem for evaluation. It makes unobtrusive measures in evaluation far more important than they are in other areas because of the hypersensitivity to the intrusiveness of the measuring instrument. There is a therapy problem of how best to handle valuephobia that we need to be more explicit about.

I want to finish with the two topics that I think are most important of all. One is a change in the basic categories of evaluation. The other relates to what are now just slogans: evaluation as a model for the social sciences (rather than the other way around) and evaluation as a new discipline.

Evaluation is normally seen as consisting of two different categories: summative and formative evaluation. It is often thought that summative evaluation is typified by a grade on a student's paper or term performance, a macro-grading. Formative evaluation, it is thought, has to be micro-analytical. It has to say: "You should improve in this way" or "if you did this, you would get better results." In other words, you try to improve the product while it is still in a flux.

Now I am going to distinguish the two components in each of those two types of evaluation. First, I will distinguish between holistic, or macro-evaluation and analytical, or micro-evaluation. Holistic evaluation is a single overall summary of the merits of something, often but not necessarily, brief. Analytical evaluation, on the other hand, is a set of sub-evaluations that are combined into an overall evaluation. There are two main types of analytical evaluation: component evaluation and dimensional evaluation. The first concerns the analysis of parts, and the second involves studying aspects of the total entity and/or of its performance. Neither gives causes or remedies, though some may be inferred.

I've taken the terms holistic and analytical from the language used in the evaluation of compositions. Holistic is the single grade. An analytical assessment involves saying that the grammar is bad, the expression is good, the character definition is bad, and so on. This is dimensional analysis. Similarly, in the evaluation of television sets in ***Consumer Reports,*** they will say that the top-rated set is a Zenith 25-inch, which does very well on color conversion, the avoidance of fringing, local or distance reception, and so on. These are dimensions of performance. You can often give an analytical evaluation that breaks down the overall performance whether it is done summatively or formatively. This is a very important step.

In a sense, such an analytical evaluation explains, justifies, and refines the results; but it has nothing to do with causation. Causation in the television case would be saying that the reason this set did badly or this one did well was the use of fine quality resistors, good circuit design, or a good picture tube. This is component evaluation plus a causal theory. These are all independent; remedies are sometimes derived from causal analysis, but often not. Folk medicine is a classic example. In that area, we find things that work, and therefore, we have remedial recommendations under certain conditions. But we don't have the faintest idea why they work, because we have no theory at all. Causes are well below the full theoretical level; they may spin off from some theories, but you can have causes without having theories. Or you may have remedy without having a theory, and without a causal account of the process, and so on. You frequently find that the Hawthorne effect in education provides remedies, but you don't know what caused the trouble. It is increasingly important to separate remedial and causal factors, because we have forced the educational researchers' perspective on evaluators for too long. This view is that you can't have remedies without causation and, in general, a causal theory. That's not true; you can.

It is normally said that teacher evaluation never should be done purely summatively, and that you must have evaluation for faculty development as well as for personnel decisions. This means giving suggestions for improvement. This is wrong. It may be the case that you lack the legitimating science to make the improvement recommendations, but you have the validity for the summative evaluation. You mustn't act as if somehow God is going to make it possible for you to tell people how to fix every error condition there is; it is not that way. In teacher evaluation we have wrongly acted as if in order to do summative, you have to do remedial. Summative evaluation, done early as holistic evaluation, is a very good basis for formative evaluation, based on trial and error. It is a professional's job to improve if a valid evaluation indicates the need for improvement. It's not an administrator's job *per se* to teach professional teachers how to improve. Professional teaching should involve improvement skills; whether it does or not is another question. But at the moment, we have to face the fact that we lack the science to support scientific remediation.

Final comment — evaluation as a model for the social sciences. We have to start thinking of evaluation, with its concern for multiple factors (such as ethics, costs, side effects, comparisons, contexts, and impact), and the failure of the hypothesis testing model as the appropriate model. Evaluating always involves "exploratory" research, and is, therefore, completely unlike the standard paradigm of social science investigation. We must recognize the fact and start changing things. We cannot continue, as we do at major schools of education, to require that the dissertation consist of five chapters, one of which presents the hypotheses to be tested. That is not evaluation. In fact, it rules out evaluation and creates the idea that the only type of research is useless research. Research based on the standard social science model is not going to allow you to say: "In my school, we have to go with this practice" or "in my classroom, we have the right to go with this improvement." Those types of recommendations require evaluation, and evaluation has to look at six, seven, or ten dimensions of a practice and its side effects, costs, and political ambience before you can make a decision. The social science model, on the other hand, is going to look at only one. That's not enough to provide connections to practice. If we want educational research to be tied to practice, we have to do better than that, and evaluation ought to show us the way.

THE PROGRESS OF EDUCATIONAL EVALUATION: ROUNDING THE FIRST BENDS IN THE RIVER

by Nick L. Smith

We have been asked in these four papers to comment on the status of educational evaluation, to address the questions, "What have been the key developments so far?" and "What are the needs for future growth?" This is a bit like asking someone who has been rafting down a churning river all morning how the trip is going. We haven't been at this evaluation business very long, and we are not terribly sure whether we are adroitly navigating each new turn or merely trying to keep upright through the changing economic currents, the encroaching banks of legislative and judicial control, and the shoals of special interest groups. We have had to gain our balance quickly, having been launched in the swift currents of the ESEA legislation in the mid-sixties. While it does seem useful to consider for a moment where we in evaluation are likely to journey over the next several years, our experience is short and our perspective does not extend much beyond the next bend.

Evaluation: Secure or Insecure

Boulding's (1980) distinction between secure and insecure science has been suggested as a common theme for each of the four papers. Boulding points out that each discipline has "more secure" knowledge or images that are well-supported and not likely to change much over time. Other elements of the discipline are "less secure" and subject to change. Overall, how would we characterize our understanding of educational evaluation? Is our knowledge more or less secure?

Following Boulding's analysis, we would have to say that our knowledge is insecure to the extent that (a) it is based on a small and possibly biased sample of phenomena; (b) it covers only a small area of a complex field; (c) it covers a short span of time; and (d) it is influenced by rare, non-reproducible events. Given these criteria, we have to conclude that our knowledge of evaluation is far from secure.

Our knowledge is based on a small, biased sample in that our attention has been focused almost exclusively on the evaluation of educational **programs** within local district settings. What is even more restrictive is that we have paid relatively little attention to anything but federally funded educational programs. The evaluation of educational personnel, materials and products, and broad systems has definitely not received as much support and interest as program evaluation. Our understanding of evaluation is further biased in that we have tended to focus on public rather than private education, basic and compensatory rather than special or supplemental education, and elementary and secondary versus preschool or higher and continuing education. I am not arguing that these areas necessarily deserve greater attention; I am only pointing out that our collective experience is more limited in these areas.

To date, we have also covered only a small area of a large and complex

field. Educational evaluation has been narrowly viewed by many to be concerned only with the use of cognitive knowledge in the improvement of formal decisions made in specific organizational settings. Only recently has that view been broadened to include such issues as the role of political and economic forces in the decisionmaking and the non-decisionmaking uses of evaluation. Still on the horizon are such concerns as the impact of personal influence and historical trends in educational decisionmaking, the relationships between values and fact in education, and alternative means of valuing.

Certainly we would all agree that our span of experience with evaluation is extremely short; and, consequently, our knowledge of evaluation is insecure from this perspective.

Is our understanding of evaluation based on rare events? Was the popular concern with the quality of education that was launched with the Sputnik a rare event? It provided the inspiration, as well as the resources, for the early work in educational evaluation. If it was a rare occasion, was it a fortunate occurrence that beneficently formed the nature of evaluation, or was the evaluation of the 1970's misshaped by a social artifact of the late 1960's? It is obviously too soon to answer these questions from a historical perspective. It will take some time before we can assess whether evaluation as it was practiced in the 1970's charted the main course for the future of evaluation or if it provided only a short trip down a minor tributary.

With so little experience to draw on, it is difficult to definitively state how secure our knowledge about educational evaluation is. I believe, however, that all indicators suggest that our understanding is highly insecure. We can expect major shifts in emphasis, expanding horizons, and radical alternatives to emerge as educational evaluation continues in its current period of expansive growth. Like shooting the rapids, living through such turmoil will result in greater security in the knowledge gained upon reaching the other side.

Evaluation: Past and Present

What have we accomplished in evaluation? Where do we stand today? If we continue the historical perspective suggested by Boulding, then we must also ask whether an enterprise only fifteen years old really has a past. For such a short period as this, the first two questions can be answered by a statement of the field's present status.

The other three papers in this collection are written by men in an ideal position to comment on the evaluation trip in education thus far. Drs. Popham, Scriven, and Stufflebeam are part of that small group of craftsmen who built the raft on which we are all now traveling downstream (or is it up the river?). I am only a recent addition to the crew and am interested in their assessment of where we have been. It is rather like they are the Samuel Pepys of evaluation history.

Before moving on to the topic to which I want to devote most of my attention, the future of evaluation, I would like to share my perception of where we presently are in this field. During the past fifteen years there has

been a lot of effort put into thinking about evaluation: what it is, what it can be, and what it should do. Much of this intellectual activity can be traced through the articles on the so-called evaluation models: the countenance model, the discrepancy model, the CIPP model, the goal-free model, and so on. These papers helped us to develop alternative views of evaluation, but were less helpful in showing us how to do evaluations.

Doing evaluations has become a big business in the last fifteen years. The number of evaluators has grown dramatically as federal, state, and local resources have been increasingly diverted to this new form of management aid. The perception of evaluation as the decisionmaker's best friend is reflected in the management orientation given to most evaluations in education. In the local schools, evaluation activities serve administrative rather than instructional needs, are designed to aid formal decisionmaking, are goal or objectives based, and are heavily dependent on measurement and testing procedures (Lyon, et al. 1978).

In the last few years, an increasing number of individuals have begun to offer alternatives to this "traditional" view of educational evaluation by advocating naturalistic (Guba, 1978), responsive (Stake, 1975), or critical (Eisner, 1975) approaches to evaluation. Evaluators have also come to view themselves more formally in recent years. The evidence of this crystallization of a professional identity is readily apparent: we now have evaluation journals, evaluation societies, evaluation standards, and various types of evaluation awards. From a part-time activity for educational researchers in the mid-1960's, we seem to have developed a new professional identity for the 1980's.

Evaluation: The Future

There are three questions one might ask about the future of evaluation: What **will** happen, what **should** happen, and what **can** happen.

Answering the question of what **will** happen requires either a crystal ball or expertise in the techniques of forecasting, depending on your epistemology. Since I have neither a crystal ball nor facility for forecasting, I am left with the expedient strategy of making guesses based on the assumption that yesterday portends tomorrow.

I anticipate that evaluations conducted within local school districts and state departments of education will continue to be heavily management focused with major emphasis on the use of testing to inform administrative decisions. There also will be an expanding array of alternative approaches to evaluation. While they will capture much attention, these new approaches will be used only in an increasing number of specialized applications and will not replace the current dominant approach.

One of the more exciting changes in evaluation will be that it will continue to become increasingly interdisciplinary. There is already a concerted movement, evident in a few evaluation journals and societies, to combine many social science disciplines (psychology, sociology, economics, political science, history, law, anthropology, and others) with most of the areas of evaluation practice (education, criminal justice, health, social

service, industry, the military, private foundations and federal oversight) to form a full-spectrum view of evaluation theory and practice. This increasing catholicity will greatly enrich the thinking on evaluation and will promote the sharing of methods and perspectives across diverse areas of evaluation practice. In fact, this movement is as likely to have a hybridizing influence on the disciplines of social science as it is to enrich the theory and practice of evaluation. If we are exceedingly fortunate, this cross-fertilization will result in more of those rare individuals who can make significant contributions across a number of disciplines, such as Michael Scriven and Donald Campbell. Undoubtedly, this diversification of evaluation will be a mixed blessing to the evaluation practitioner who will be faced with an expanding array of methodological options and potential social roles.

Of course, many of evaluation's current issues will continue to be of concern; we will continue to deal with:

- improving the relevance and utility of evaluation studies,
- increasing the role of evaluations in providing equal educational opportunity,
- improving the cost-efficiency of evaluation efforts,
- expanding the role of consumers and stakeholders in evaluative activities,
- integrating evaluation effectively within legislative, judicial, and administrative functions,
- making major advances in the "assessing value" side of evaluation,
- strengthening the evaluator's ability to deal with moral and ethical problems, and
- creating a professional identity or identities to guide the future of evaluation.

How would we answer the question of what **should** happen in evaluation? I suppose one could proffer a reasoned argument based on an empirical needs assessment of the profession. I do not believe, however, that professions, or any social organizations, for that matter, follow such a rational model; their directions are dictated by broad social forces and the energetic activity of a small number of influential leaders. The currents of the river and the hard work of the few individuals holding the paddles are what determine whether we negotiate the next turn successfully, capsize, or beach on a shoal. Therefore, I find personal statements of what direction evaluation should take to have more significance than generalized arguments of desirable conditions.

Rather than list a number of activities that I believe would contribute to evaluation theory and practice in the coming years, I will emphasize the one area that I think can make the greatest contribution: the conceptual and empirical study of evaluation practice. We need research on evaluation; we especially need grounded, empirical studies of evaluation practice. We have almost no descriptive information on the practice of evaluation, few field studies of evaluation impact, and scant attention to the empirical study of evaluation method. I have discussed the need for this type of work

elsewhere (Smith, 1979, 1980(a), 1980(b)) and will not belabor the point here, except to outline the focus of this research.

Research on evaluation should focus on (a) understanding the nature of the **contexts** within which evaluation takes place; (b) providing information on the effectiveness of the various evaluation **methods**; (c) improving the **utilization** of evaluation results; and (d) explicating the empirical connections between the evaluation problem, local context, evaluation method, and evaluation utilization. With such background research, it would be possible to begin developing a theory of evaluation practice. This long-term activity of providing an empirical basis for evaluation practice would do the most, I believe, to improve the profession of educational evaluation.

The third question concerning the future of evaluation is what **can** happen. This question calls for a kind of vision infrequently practiced in evaluation, or in science, for that matter. Loren Eiseley (1973) describes Frances Bacon as one of the few individuals with this capacity. In his book on Bacon entitled ***The Man Who Saw Through Time,*** Eiseley says,

> Bacon was the first great statesman of science. He saw its potentiality in the schools; he saw the necessity of multiplying researchers, establishing the continuity of the scientific tradition, and promoting government supported research for those studies which lay beyond private means and which could not be accomplished "in the hourglass of one man's life." (1973, p. 36)

And Bacon saw this over 400 years ago!

We can attempt to follow Bacon's example of seeing what is possible in the future, noting, however, a comment made by archbishop Richard Whately, "Bacon's wisdom is like the seven-league boots which would fit the giant or the dwarf, except only that the dwarf cannot take the same stride in them" (Eiseley, 1973, p. 47). You must measure for yourself the length of our visionary strides into the future.

To provide a glimpse into the future of what is possible in evaluation, I have prepared five vignettes, which might be called "scenarios of the possible." * These scenarios describe possible new methods. I have extrapolated from some of the emerging new approaches in evaluation to generate these fantasies; but in order to give them reality, I wrote these scenarios as if the methods were now fully developed. These are clearly fantasies at present, but they do suggest what could happen in evaluation; and as Boulding notes, "...fantasy is part of the appropriate methodologies of all fields of human knowledge" (1973, p. 836).

The Investigative Approach

Jonathan, in the outer office of the State Superintendent of Public Instruction, waited impatiently with the results of his latest investigation. The Superintendent had to be made aware of what Jonathan had found so that he could take action before either the legislature or the newspapers

*My thanks to Adrianne Bank for suggesting this term.

learned what was going on.

Jonathan worked as a special assistant to the Superintendent. He did in-house evaluative studies, usually following up leads of illegal, unethical, or incompetent behavior within the state education system. Jonathan's position had been created by the Superintendent, who was elected during a period of reduced public support and funding for the state educational program on the campaign promise that he would clean up corruption and waste within the educational system. Jonathan's reports were frequently politically sensitive, and his position had to be protected by being attached directly to the State Superintendent.

Jonathan thought back over the latest investigation. As usual, he had attempted to reconstruct a sequence of events through the review of extensive documentation and by interviewing key informants. He had been able to establish a plausible chain of events including effects, causes, and participant intents. Jonathan's methods have improved over the last several years of practice, but they basically grew out of the approaches used by investigative journalists, congressional investigative teams, criminologists, and field researchers in the social sciences.

Jonathan usually provided his services when quick studies were needed; he worked in naturalistic settings under considerable pressure. While he often did short review studies as a means of providing information to the Superintendent when he was drafting new policy, Jonathan worked mostly on uncovering fraud or incompetence. He relied on existing information sources, collecting data primarily through interview, observation, and document review. Most recently he had conducted evaluative investigations of the state materials procurement procedures (there had been charges of kick-backs), of sexist discrimination in the hiring of elementary school principals, and of charges that certain professional schools had obtained copies of the state certification tests and were giving their students an unfair advantage in passing the state professional exams.

Jonathan's methods provided one of the few evaluation approaches which allowed for quick field studies. Nevertheless, his investigations frequently established only strong **prima facie** cases and did not provide incontrovertible evidence. Since the objects of his investigations were often highly controversial, his was a stressful position.

Jonathan reconsidered the details of his current investigation, double-checking that he had pursued every avenue and had built a report that would stand up to public scrutiny in the newspapers, and, if necessary, in the courts.

The Hearings Approach

Mario slowed his pace as he crossed the street. He did not want to be too early for the meeting of the Citizen's Advisory Committee. It was always difficult to keep the committee's attention on the prepared agenda when he arrived early and became embroiled in a discussion of the latest political issues in the case. Mario worked for a private foundation, although much of the resources for his evaluative work came from state and federal

government subsidies and from citizen contributions. It was Mario's job to help minority and special interest groups conduct public evaluations of broad issues concerning the public interest. In this position, he acted as master methodologist, as non-partisan chairman, and occasionally as individual counselor. As he continued down the street, Mario reflected on how much more involvement and influence minority groups now had in education as a result of the types of public evaluations he and others were helping to create.

The committees of specialists and lay public, which Mario organized and worked with, gathered evidence about the multiple arguments surrounding an issue of public concern. They then employed public hearings to scrutinize and, ideally, resolve major points of disagreement. Although there were now several variants of this basic approach, the model derived originally from the hearing procedures used in Congress, legal trials, public environmental impact studies, and national investigative panels. This approach was increasingly being used not only to heighten the participation of various minority groups in education, but also when public values, citizen input, and the analysis of educational problems were seen as more crucial than the collection and analysis of technical data. Relevant data, arguments, and position statements were prepared and presented under various types of hearing formats.

Mario had a diverse methodological background, since it was important for him to be able to summarize multiple types of data (such as research findings, case studies, and anecdotal information) within coherent position statements. His training and special interest in public activism were also important, since he had to conduct public hearings so that useful analyses and closure could be reached. Because of the frequently sensitive topics being evaluated, Mario also had to be politically astute.

Approaching the building where the meeting was to take place, Mario reassessed the progress of the current evaluation. The citizen committee was making good progress; it appeared that the hearings scheduled for next month would allow for a sharing of diverse viewpoints on the problem. Unfortunately, as was usual with this approach, some complex issues were being overly simplified, and the process was being subjected to political influence by certain groups. Nevertheless, having presented the State Board of Education with the anticipated roster of witnesses who would be making prepared statements at the hearing, Mario felt confident that this study would have some influence in shaping the state's policy in the area.

The Criticism Approach

Su-Lin sat in the staff lounge of a small elementary school reviewing the lesson plans of several teachers. She was one of the better-known evaluation critics in this region and was currently involved in a study of experiential materials used in values education. As a private consultant, she had done a series of independent evaluations of educational programs. In a number of cases she had been hired by state and local agencies to

perform external evaluation critiques of materials and programs.

Following her usual strategy, Su-Lin was reviewing materials prior to doing some classroom observation. Drawing on her extensive experience with similar projects, she would then conceptually analyze the materials under study to assess their quality and to illustrate how their constituent elements contributed to that quality. While most critics specialized in product evaluation, Su-Lin had done some work in program evaluation as well. Her primary focus was on the program's or product's quality and its relation to current public needs and interests. Using techniques of critical anlaysis, she derived insights about how the product's or program's form and content contributed to its function. She then shared these insights with her clients.

This criticism approach to evaluation had developed through the adaptation of criticism methods used in the fields of poetry, art, music, film, and architecture. The approach also had been significantly influenced by consumer advocate interests in evaluation and had been used in studies of materials in basic skill areas, such as math, reading, and computer use. Su-Lin had just completed a study of classroom climate and student-teacher relationships in several vocational schools.

Su-Lin's training had involved extensive prior experience with the types of materials she routinely evaluated. Her effective written communication skills were also essential in her work, as was her intimate understanding of the techniques of conceptual analysis and of the history of professional criticism in her particular area of expertise.

Su-Lin's work involved the use of seasoned, experienced judgments of quality and was one of the few approaches which allowed one to judge a current product in light of a large number of similar previous efforts. Her best insightful critiques had done much to illuminate those elements of various products which contributed to their quality. Unfortunately, the criteria and experience employed in making these judgments usually could not be made public because they resided within Su-Lin; and so to some extent, her clients had to validate her professional judgments through their own experience with similar products.

As Su-Lin left the staff lounge and walked down the hall, she was greeted by two passing teachers who immediately recognized her, since she was one of a small number of regionally-known educational critics. The children filed into the classrooms. She quietly took a seat in the rear to begin her observation of this class, which was like hundreds she had seen in the past, but was also in its own way unique.

The Optimization Approach

Having refilled her coffee cup, Katherine returned to her desk to study the latest computer printouts. Her current study concerned an evaluation of the allocation of remedial mathematics materials across the thirty-seven schools in the large school district for which she worked. Katherine's work in evaluation required high technology as well as considerable expertise and material resources. Many other evaluators doing the kind of work she

did were employed in large state and governmental agencies and in the largest school districts.

As usual, Katherine was attempting to model the system under study to analytically assess its parameters and to maximize its functioning. Her work involved the evaluation of complex systems where the optimization of operations, resources, and impact were of primary importance. Modeling procedures and computerized mathematical analysis were her standard tools. Focusing primarily on concrete, quantifiable parameters, her work was highly analytic and oriented toward better control of the system.

Although the approach Katherine used in educational evaluation had been employed for many years now, its history was even longer. It had been derived from similar procedures used in operations research, cost-benefit analysis, management control theory, and general systems analysis. She had had extensive academic preparation in mathematical modeling, computer analysis, and optimization techniques. Understanding the design of human systems was also a significant element in her training. As was common, most of her studies concerned the optimization of educational subsystems; for example, work in facilities, transportation, student counseling, and some work in instructional design.

Katherine found the figures she had been looking for. She was always pleased when the modeling procedures resulted in an efficient redesign of system elements. Being able to quantify and optimize system operations gave her pleasure; however, she continually had difficulty with the nonquantified factors, such as personal values, student morale, and the nonrational influences in human affairs. Her use of the strong analytic tools was especially effective when system control and optimization had been a primary educational goal. Preparing the data for the next round of simulations, Katherine thought how well this particular study seemed to be going.

The Representation Approach

Billy sat in his study at home preparing notes to teach his university class on evaluation the day after tomorrow. Most of his presentation would concern his recent evaluation of a summer film institute. Using the representational approach to evaluation, which was highly favored by people associated with arts education, he spent much of his time evaluating film workshops, art and museum exhibits, plays, and dance theater. His approach was to capture the experiences of participants in these activities, to analyze those experiences in an appreciative sense, and then to construct a surrogate experience so that nonparticipants could make their own assessments of the quality of the activities' educational and artistic offerings. Thus, his work most often resulted in evaluations that were shared experiences rather than written reports.

Billy had arrived at this approach to evaluation only recently, spurred by his interest in representing a particular educational program or event to an external audience. He had been successful at using a variety of techniques

to recreate educational experiences, including photographs, mime, plays, musical compositions, paintings and stories; and he had used these to share an educational experience with audiences, so that they could make th ir assessments of its quality and worth. This approach to evaluation had developed from a combination of procedures used in art, music, theater, and case study work.

Billy's training had required him to increase his understanding of how an experience could be altered through various forms of representation. He had developed a sensitivity to participant reactions and had acquired good observational and interview skills. What he liked most about the approach to evaluation, however, was that it required much creativity in producing a first-hand, high-fidelity surrogate experience that enabled persons in the audience to make their own assessments of the educational program's quality. Most of these evaluations were conducted by experts in some area of the arts, and Billy was pleased to be a part of that group.

Billy continued preparing his class notes, outlining the major po nts he wished to raise. This approach to evaluation was one of the few models suitable for use in the arts—the criticism approach being another. In contrast to the criticism approach, however, which provides an external assessment of a program's or event's quality, the representational approach depends heavily upon the fidelity of a surrogate experience and on the audience members to make their own assessments of quality. Of course, these evaluations tend to be non-replicable and are usually of greatest utility to people who are already familiar with events like the one being evaluated. Billy wanted to emphasize to his students that this was a specialized form of evaluation; one which most involved the clients in making their own assessments of quality, and which concurrently increased the clients' capacity for doing more of their own evaluations.

Do these scenarios help us envision what can happen in evaluation? I think they do begin to suggest alternatives and options that are only dimly perceived at present. They are not much like the pre-post test/experimental design approach that is currently dominant in educational evaluation; but then some radical alternatives might prove very useful. After all, part of completing a successful rafting trip depends on knowing when to carry the raft around the rapids rather than shooting through them.

We have only just started our evaluation trip. The current is swift, and we have only made the first few turns in the river, so our vision is limited. Perhaps considering visions of what is possible ahead will help us deal with the future of educational evaluation as it unfolds before us.

REFERENCES

Boulding, Kenneth E. Science: Our common heritage. ***Science, 207*** (4433), 1980, 831-836.

Eiseley, Loren. ***The man who saw through time.*** Charles Scribner's Sons, New York, 1973.

Eisner, Elliot W., ***The perceptive eye: Toward the reformation of educational evaluation.*** Stanford Evaluation Consortium, Stanford University, California, December, 1975. (ED 128408; TM 005591).

Guba, Egon G., ***Toward a methodology of naturalistic inquiry in educational evaluation.*** Los Angeles: Center for the Study of Evaluation, University of California, 1978.

Lyon, Catherine D.; Doscher, Lynn; McGranahan, Pamela; & Williams, Richard. ***Evaluation and school districts.*** Los Angeles: Center for the Study of Evaluation, University of California, 1978.

Smith, Nick L. Requirements for a discipline of evaluation. ***Studies in Educational Evaluation, 5,*** 1979, 5-12.

Smith, Nick L., Studying evaluation assumptions, ***Evaluation News,*** No. 14, Winter, 1980, 39-40. (a)

Smith, Nick L. ***Evaluating evaluation methods.*** Paper presented at the annual meeting of the American Educational Research Association, Boston, April, 1980. (b)

Stake, Robert E. (Ed.) ***Evaluating the arts in education: A responsive approach.*** Charles E. Merrill, Columbus, Ohio, 1975.

CRUMBLING CONCEPTIONS OF EDUCATIONAL TESTING

by W. James Popham

When I was a graduate student in the late fifties, I first encountered the writing of anthropologist Ralph Linton. I was particularly taken with Linton's analysis of culture, in which he observed that a society's stability derived chiefly from the extent to which elements within the society's culture were generally accepted by its members. Other writers, employing Linton's analysis of culture, noted that the core of societally accepted cultural elements in mid-century America had dramatically diminished; this, according to them, was chiefly because of technological innovations.

As we consider the field of educational evaluation, now in its second decade as a formal discipline, I am impressed with the relevance of Linton's analytic structure for those of us specializing in evaluation; this is particularly true with respect to the realm of educational testing.

When we became serious about the enterprise of educational evaluation, sometime in the mid-sixties, we encountered a number of well-established expectations regarding educational measurement. Educational tests had been prominently used in American education since the twenties; and United States educators carried around a set of widely-held views regarding what tests were supposed to look like, how they were to be used, and what kinds of interpretations of pupils' test performances were appropriate.

Thus, when individuals began dabbling in educational evaluation shortly after the enactment of the Elementary and Secondary Education Act (ESEA) of 1965, they quite naturally carried those solidly entrenched conceptions of measurement with them. In the late sixties, of course, there were very few educational evaluators who had been trained specifically for that assignment. Most of us were retreads from other fields, typically research and measurement, since it was thought in those days that educational evaluation was little more than warmed-over research with an applied focus.

But traditional trappings often inhibit those working in untraditional roles. And such was the case with educational evaluators' conceptions of educational testing. *For more than a decade, educational evaluators have gone into battle burdened by outmoded and counterproductive conceptions of educational measurement.* In many instances these traditional views regarding measurement have proved to be so pivotally dysfunctional that an entire evaluation endeavor has yielded little, if any, useful information.

It is not the case, obviously, that all pre-sixties advances in educational measurement are to be repudiated. Much of that pre-ESEA technical base in measurement is thoroughly appropriate; but much of it isn't. A challenge of the eighties for educational evaluators will be to scrutinize the field of educational measurement, now in a state of Heraclitian flux, to sort out that which sensibly pertains to the tasks of the educational evaluator.

In the remainder of this analysis I shall set forth some time-honored conceptions of educational testing; conceptions that, though widely-held,

are deservedly beginning to collapse. For educational evaluators to proceed as though these views of measurement will or should be sustained is foolhardy. Today's educational testing world is different from the one in which we were reared. Educational evaluators must live in the present.

Crumbling Conception One: Because of its self-evident dividends, educational testing will be viewed as acceptable by classroom teachers and school administrators.

Educational testing has been part of our public school practices for so many years that we assume teachers will regard it as an annoying, but nonetheless essential, ingredient of the overall educational enterprise. When the district superintendent fires off a memorandum to principals and teachers calling for six hours of standardized tests in connection with Title I funding, or for four hours of testing for the statewide assessment program, it has been assumed that teachers and principals will docilely, if not jubilantly, comply.

Legislators have historically viewed taxpayers in an analogous fashion; that is, "They may squawk, but they'll pay up!" But that perception of taxpayer acquiescence blew apart with the passage of tax-limitation initiatives in the late seventies. Too many taxpayers believed themselves pushed beyond their limits. Perceived excesses led to taxpayer revolts.

In this spring of 1980, a bill has been introduced in the California Senate to prohibit state regulations which require that "more than one standardized achievement test be administered annually per student." This legislative proposal springs chiefly from the initiatives of state teachers' organizations. Leaders in these organizations believe that teachers, like taxpayers, have been pushed beyond their limits with testing programs that provide little useful information for them. Anti-testing revolts may soon be upon us.

I shall not undertake an appraisal of the merits of these anti-testing reactions, but it is easy to understand how classroom teachers would recoil from the apparently relentless and ever-expanding intrusions on their instructional programs. And despite test publishers' reassurances about the "instructional utility of test results," the bulk of teachers can't seem to find any substantive instructional dividends from most of these testing extravaganzas.

The point of such developments is simple. For educational evaluators to assume that in the future they will be able to count on yesteryear's testing programs is naïve. We may soon find more than one state where **by law** children cannot be tested more than a specified number of times annually. Under such circumstances, to design an evaluation study which calls for the administration of additional tests would be absurd. Yet, to rely on evidence from regularly administered, but off-target, tests would be equally inappropriate. The problem is nontrivial.

Evaluators may merely adopt a passive stance on this issue, awaiting the results of legislative struggles between teachers' organizations and the

testing industry. But, since evidence of pupil performance is such a crucial component of most evaluation studies, it would seem more appropriate for evaluators to enter this fray. An enlightened stance regarding the use of educational tests is possible. It will not be the traditional "Tests are good for you, teachers" platitudes of the past or today's battlecry of "Toss the tests in the paper shredder." Educational evaluators have a vital stake in seeing that this issue be resolved so that, **although the interests of teachers' organizations and testing companies are considered, the interests of pupils are paramount.** A moderate "test only when appropriate" policy should be promoted quickly or the examiners will be legally barred from the examinees.

Crumbling Conception Two: Achievement tests should mirror classroom instructional emphases.

For years we prepared standardized achievement tests in such a way that they would, insofar as possible, parallel what was going on in the schools. It was impossible to pull this off with precision, since standardized tests attempt to serve a whole nation; however, efforts were made to create achievement tests that took as much cognizance as possible of curricular emphases throughout the nation. The most widely used textbook series were surveyed assiduously by test-makers, since it was recognized how significant such textbooks were in determining what goes on in classrooms. If the tests reflected what was in the textbooks, it was thought, then the tests would do a reasonable job in reflecting what went on in classrooms. The logic was flawless.

But in recent years we have seen test results taking on a far more formidable function. As never before, students' test performances are being used as indices of program quality; that is, as indicators of the effectiveness of teachers, schools, school districts, and even state education systems. Because the stakes associated with the attainment of good test scores have risen so dramatically, more and more educators have found themselves pondering the tests themselves as instructional targets.

Most American teachers have been admonished *ad nauseum* of the perils inherent in "teaching to the test." Some teachers-in-training, perhaps not too conversant with formal religion, no doubt assumed that "Thou shalt not teach to the test" was one of the top ten pedagogical commandments.

But in recognition of the powerful contingencies that have become hooked up with test results, such as next year's state and federal funding, some educators are beginning to re-think the efficacy of an instructional stance that tells teachers to blindly teach to the very tests that will be employed to assess the quality of that teaching.

Stuart C. Rankin, a key administrator in the Detroit public school system, now embraces the notion that *properly fashioned tests can and should become the force that drives curriculum.* Rankin and his Detroit colleagues have installed a competency testing system which unabashedly

attempts to stake out testable competencies in an effort to clarify instructional expectations. By more sharply explicating the desired consequences of instruction, and by urging Detroit teachers to provide on-target instruction, Rankin and other Detroit educators hope to bring about tangible boosts in the effectiveness of Detroit teachers. Those teachers are urged to teach to the test, not in the sense of a particular set of test items, but in the sense that the test represents defensible competencies which students should acquire. The Detroit motto is "provide direct instruction for each competency deemed important."

Conceiving of tests as instruments for determining the emphases of an instructional program, rather than as sponges that only soak up its effects, is a new and potentially powerful way to view educational tests. Indeed, since many competency tests carry with them important contingencies, for instance, whether or not a student receives a high school diploma, such **competency tests will serve as curricular magnets.** Today's educational evaluator must recognize that the former notion of tests as only curricular mirrors is decisively out of date. Educational tests will now serve in some cases as **interventions,** not indicators.

Crumbling Conception Three: Educational tests provide equitable assessment opportunities for students of differing races and socioeconomic backgrounds.

Even though for the past decade or so we have heard a fair amount of rhetoric tossed around regarding test bias, there is precious little evidence to indicate that test-makers have dramatically altered their practices in this regard. Furthermore, if one considers the amount of solid empirical work that has been carried out in an effort to isolate, hence eliminate, causes of test bias, we might well argue that the eradication of test bias constitutes an essentially unstudied problem.

Even though test publishers are doubtlessly aware of the possibilities that their tests may be biased against minorities and the poor, there are few truly exemplary test development projects in which intense efforts to eliminate test biases have been present.

In a recently published reading achievement test, for example, an all-too-common instance can be seen of how publishers cope with the problem of test bias. For this test, whose cover is adorned with drawings of children clearly reflecting varied races, we learn that all items were independently reviewed by a team of three bias reviewers: one white, one black, and one Hispanic. Any item was rejected which was judged to be biased by **all three** reviewers. Clearly, the bias review standards for this test were somewhat short of stringent.

Yet, in spite of what is surely only superficial treatment of the test bias issue, both with respect to empirical study and to the implementation of anti-bias test development procedures, educational evaluators often blithely assume that discriminatory testing practices have been eradicated. It's almost as if we want to believe that once a problem has been isolated

and chattered about, someone must have taken care of it. In the case of test bias, the rhetoric is there but the remedies aren't.

Crumbling Conception Four: Educational test development agencies are benign, quasi-scholarly organizations engaged chiefly in research and development regarding the measurement process.

When I was a graduate student, I relied heavily in my measurement courses on the various technical memoranda dispensed by the major test development agencies. These documents were first-rate analyses of various psychometric problems then vexing the testing community. As was the case with most of my fellow graduate students, I formed a thoroughly positive perception of test publishers. The folks operating those organizations, or so it seemed to me, were equal partners in the programmatic research development to learn more about how to make educational testing more defensible.

But graduate students grow up—slowly. Begrudgingly, I now must admit that almost all test development organizations that I know are **businesses.** As such, they are dominantly guided by the profit motive. The people working for test development firms, of course, are not thereby rendered malevolent shysters. Some of my best friends work for test publishers (they keep our friendship quiet). On the other hand, some of my best friends are malevolent shysters.

Educational evaluators must view with realism the current state of educational measurement. The vast majority of tests the evaluators use are churned out by companies whose major mission is to satisfy their corporate stockholders. In the long run, we will hopefully find that those companies exercising concern for rigor and fairness—in contrast to those searching for a quick profit—will emerge as the most successful. Yet it would be folly to think that test development firms are preoccupied with improving test quality, even though fiscal losses rather than profits will ensue. That's just not the way the test publishing industry operates.

Educational testing in the United States is big, big business. And the decision rules used in big, big business have historically failed to place the consumer's interests at the top of the priority ladder. When using commercially developed tests, let the educational evaluator beware.

Crumbling Conception Five: Decisions regarding the technical qualities of educational tests will be made by psychometrically knowledgeable educators.

How valid is this test? How reliable is it? Is its standard error of measurement sufficiently small? If such questions had been asked a decade ago, we would search for answers by rushing to the test's technical manual and locating the relevant correlation coefficients and other statistical indicators. Everyone knew that to answer technical questions regarding test quality, one deferred to the psychometrically anointed. It was the

psychometricians who created the technical indices of test quality, and it was the psychometrically knowledgeable who could understand such indices. These would be the folks who would, and should, respond to the question of whether a test was technically good enough.

In recent years, however, the rules have changed. We find groups of parents initiating class action suits against state departments of education, charging that a state's "functional literacy" test is not valid, and hence should be abandoned. Another class action suit charges that group intelligence tests penalize minority youngsters. A married couple sues a school district on the ground that the minimum competency test that their son failed was not a sufficientiy defensible measure to preclude his receiving a high school diploma.

And in all of these instances it was the **courts,** not the psychometricians, who decided whether the tests were legitimate measures. It is true, of course, that psychometricians were called in as witnesses; but the final decisions were rendered by judges, not measurement specialists.

The implications of such events for educational evaluators are clear. No longer can we assume that a test which has won the approbation of its psychometric progenitors will automatically secure similar approval from the judiciary. The standards for gauging test quality will have to be made considerably more stringent in view of the lurking courtroom barriers that tests may be obliged to surmount.

In the old days, the public was docile and deferent in the face of technical expertise of any sort, and that included the field of educational testing. Citizens assumed that they could not fathom the esoterics of sophisticated test construction. After all, what normal person would not flinch a bit in the face of a factor analysis with varimax-rotated vectors?

But the public no longer genuflects in the presence of measurement experts. Special interest groups, such as organizations of minority citizens, are questioning the technical adequacy of the tests which continue to place too many of their members at a disadvantage. Such public vigilance regarding educational tests is, in my view, long overdue. The era of "snowing 'em with statistics" is at an end.

Educational evaluators, therefore, need to prepare for an onslaught against testing instruments that cannot be dismissed readily by referring the matter to the technical wizards. Educational tests must be of sufficiently high quality to be defended in language comprehensible to the average enlightened citizen or jurist.

Amid the Crumbles

With so many heretofore stable truths in educational testing beginning to totter, one might be inclined to call for an end to educational measurement. After all, if educational testing is so fraught with frailties and so subject to political, economic, and judicial perils, why test at all?

Well, albeit far less than perfect tools, educational tests do provide key information which helps educational evaluators more defensibly assess the merits of educational phenomena. If the quality of those evaluations can

be enhanced, even by a small percentage, by relying on student test results, then we must covet such test data.

Although today's educational evaluators are dealing with a new world of testing, a world far more complex than in former years, it is a world in which we still must function. And, given our dwindling core of stable measurement verities, this means that evaluators must be particularly attentive to technical and conceptual advances in educational testing. Testing has become too important to leave to the testers.

STANDARDS, RESEARCH, AND TRAINING: THREE PRIORITIES FOR PROFESSIONALIZING EDUCATIONAL EVALUATION

by Daniel L. Stufflebeam

INTRODUCTION

Conferences, symposia, and publications devoted to forecasting the future of society and its various enterprises are commonplace this year. It is easy to find projections about general areas such as the economy, war and peace, science and technology, religion, and politics. There is also an abundance of forecasts for professional areas such as medicine, law, and education. And the future of specific careers within given professional areas, such as educational administration and educational evaluation, have also received much attention.

I have been exposed to and involved in a number of such forecasting exercises. My reactions are mixed. On the one hand, the topic and treatment have at times seemed ritualistic, tedious, and superficial. On the other hand, I have found that thinking about the future and attending to other persons' views about it can be a mind-expanding experience. It has especially helped me to view educational evaluation in a broader context than is my custom and to speculate on how this field might look in the 21st century. It has also helped me to think about what is right and wrong in the field and to start thinking about an agenda that might be pursued to both improve and expand the field's contributions. On balance, I believe that thinking about and discussing the future of one's profession is a constructive and potentially productive activity.

Therefore, I am glad that Wayne Welch and his colleagues at the University of Minnesota have organized this conference. I hope that our deliberations here and the conference report will help us and others to take stock of educational evaluation, both intrinsically and extrinsically; to identify areas for improvement; and to start building and pursuing a realistic agenda of professional development and service.

These hopes, of course, presume that educational evaluation is a field that should and can make unique contributions to society. These presumptions have been questioned by some educational evaluators, such as Egon Guba and Bob Stake. Dr. Welch has invited the speakers at this conference to consider whether, and if so, in what respects, educational evaluation has a secure future.

I shall begin my paper by addressing this issue. Thereafter, since on balance I will argue that educational evaluation potentially has an important and viable future, I will consider three areas that I believe can help to ensure such a future: standards, research, and training.

I chose the first area because after four years of intensive work by a joint committee (appointed by twelve national associations concerned with education), the ***Standards for Evaluations of Educational Programs, Projects and Materials*** soon will be published. In my view, the ***Standards*** constitutes a potentially powerful tool for upgrading the practice of

educational evaluation. (Hereafter, I shall refer to the 30 standards contained in the overall ***Standards*** document.)

But until validated, the 30 standards, irrespective of the care with which they were developed, are only untested hypotheses about what constitutes sound and productive evaluation. The standards thus highlight the need for research on educational evaluation; moreover, they provide a conceptual framework for the needed research. It is both timely and important to begin examining the standards in order to derive and test pertinent hypotheses. The results of such research should prove useful for promoting the sound use of the standards, for expanding and revising them, and, in general, for developing a scientific understanding of educational evaluation

In addition to research on the standards, I want to discuss evaluation training, because I believe evaluation cannot adequately serve education and society unless its specialists and clients are appropriately trained, and because I believe evaluation training borders on being a disaster. From my vantage point, evaluation training in education is poorly conceptualized, disorganized, out of step with realistic evaluation issues, superficially implemented, and grossly underfunded. While there are a few strengths on which to build, an abundance of opportunities for cost-effective evaluation training go unused and, in general, there is a dearth of leadership of any kind in educational evaluation training. Mainly, I will argue for increased attention to this area, but I will also offer some ideas for improving evaluation training.

The four topics I have chosen to discuss—the viability of educational evaluation, standards, research, and training— are, I believe, crucial considerations for designing an agenda to improve educational evaluation as a field. There are other important areas, such as organizational development, theory, and methodology development. While they are outside the scope of what I can consider here, such topics also need attention.

Viability of Educational Evaluation as a Profession

My first topic concerns the bottom-line question of whether educational evaluation is or can become a viable profession. In addressing this question, I think it is important to consider what are meant by the concepts of profession and of a secure or viable profession. Definitions of these concepts then can be used to examine the status of the educational evaluation field.

For a definition of profession, I have turned, as is my habit, to a dictionary. According to that source a profession is "a calling requiring specialized knowledge and often long and extensive preparation including instruction in skills and methods as well as in the scientific, historical, or scholarly principles underlying such skills and methods, maintaining by force of organization or concerted opinion high standards of achievement and conduct, and committing its members to continued study and to a kind of work which has for its prime purpose the rendering of a public service."

Given this definition, educational evaluation looks like an emerging profession. The methodology and literature of this field have expanded tremendously during the past 15 years. Specialization in the field usually requires graduate-level training, and several universities offer masters and doctors degress in educational evaluation. Standards will soon be published for the purpose of promoting high quality performance by evaluators. Also, there are several organizations — especially Division H of the American Educational Research Association and the Evaluation Network — which exist primarily to advance the theory and practice of educational evaluation and to provide continuing education opportunities for evaluators. Finally, thousands of persons have jobs in which they evaluate educational programs, train evaluators, and/or conduct research and development in the area of educational evaluation; many of these jobs are supported through millions of dollars in public funds.

But even if educational evaluation looks like an emergent profession, we can also ask whether it is mainly secure or insecure. Or, more pointedly, is educational evaluation only a passing fad?

According to Boulding, the critical characteristics of a secure field are an adequate historical record of the field (i.e., the field's development was deemed important enough to record); a knowledge base that covers the total field; structures and relationships within the field that are clearly defined and relatively simple; and a concern with events that are common and recurrent.

Without going into a detailed discussion, it seems clear that educational evaluation has to be rated against Boulding's factors as a relatively insecure or immature profession. Even though the history of educational evaluation as a field of professional practice is short (whether one dates it back to Tyler's pioneer efforts in the 1940's or to the evaluation requirement in the Elementary and Secondary Education Act of 1965), it is largely unrecorded. In addition, little research has been done on the practice of educational evaluation. While some important studies have been done, the base of knowledge about evaluation practices and their effects is meager. And the field hardly deals with structures and relationships that are clearly delineated and simple. On the other hand, the phenomenon of formal evaluation in education is present in some form in virtually all educational enterprises, and many aspects of it are recurrent and amenable to study.

While the field appears to be mainly insecure against Boulding's factors, some of the insecurity could be dealt with. Certainly those persons concerned with promoting and sustaining good evaluation practice would do well to systematically record and analyze the history of educational evaluation and to conduct other forms of research on the area. In encouraging the maturation of the field, scholars of educational evaluation should also extend their efforts to develop a sound theoretical base for the field.

Beyond those factors mentioned by Boulding, it is also crucial in considering formalized educational evaluation to remember that education

is a public concern. When Bob Stake argued at a symposium at our Center that educators should avoid doing evaluation, I asked him why he continued to do it. He responded that evaluation is inevitable since society will continue to require it. He added that he wants to help educators to do evaluation seldomly and inexpensively; but when it has to be done, he wants to help them do as useful research as possible. In that explanation, I believe Stake put his finger on a key reason why educational evaluation will endure as a field of professional practice. Taxpayers and consumers are acutely aware that education is vital to society and that it is expensive; consequently, they want and often demand to know whether education is efficient, effective, and accountable. The only way to satisfy these concerns, short of a con-game, is through sound evaluation. Since evaluation of education is going to be required and done, educators should conduct it so as not only to reveal but also to assure the quality of their work. I believe this is a practical reason for promoting quality evaluation in education. In addition, if educators are to lay claim to professionalism, they must assess their goals, plans, operations, and achievements, both to ensure and to demonstrate that they are helping to meet the educational needs of their students and of society. Thus, I rest my case for promoting the continued development of educational evaluation as a viable and important field of professional practice and as one that needs and merits increased attention.

The Standards

As I noted earlier, one characteristic of a profession is the maintenance of high standards of achievement and ethical conduct. To this point in the brief history of educational evaluation, the practitioners of this field have had no widely-shared, comprehensive, and carefully articulated set of principles by which to guide and assess their work.

But that circumstance is about to change. In August 1980, the McGraw-Hill Company will publish ***Standards for Evaluation of Educational Programs, Projects, and Materials.*** That book was prepared through an extensive development process that involved the efforts of about 200 people and required more than four years of work. A joint committee, whose 17 members were appointed by 12 organizations, led the effort. Currently, the committee is turning its attention to promoting the sound use of the ***Standards*** and to launching a process of ongoing review and development.

I strongly recommend that persons concerned with advancing the practice of educational evaluation obtain, study, and apply the ***Standards.*** They will find explicated in that document 30 principles which collectively promote evaluations that are useful, feasible, ethical, and technically sound. For those of you who haven't been exposed to drafts or the finished ***Standards,*** I will briefly describe its contents. I know that many persons here have helped develop and assess the ***Standards,*** and I hope that they can bear with me as I present what surely is redundant information from their perspective.

There are eight *utility* standards. They reflect the general consensus that emerged during the late 1960's concerning the need for evaluations that are responsive to the needs of clients. The concerns addressed by these standards include audience identification, evaluator credibility, information scope and selection, valuational interpretation, report clarity, report dissemination, report timeliness, and evaluation impact. In general, the utility standards require evaluators to acquaint themselves with their audiences, to ascertain the audiences' information needs, and to report the information clearly and when it is needed.

The second set of standards, that concerned with *feasibility*, is consistent with the growing realization that evaluation procedures must be cost-effective and workable in real-world settings. These standards require that an evaluation plan be operable in the setting where it is to be applied, and that no more resources and personnel time than necessary be consumed. The three specific concerns dealt with by the feasibility standards are practical procedures, political viability, and cost-effectiveness of the evaluation. Overall, the feasibility standards call for evaluations that are realistic, prudent, diplomatic, politically viable, and frugal.

The third group contains the *propriety* standards. Evaluations affect many persons in many ways, and these standards aim at ensuring that their rights will be protected. These standards especially prohibit unethical and careless actions by those who produce evaluation results. The eight concerns addressed by the propriety standards are contracting, conflict of interest, full and frank disclosure, the public's right to know, rights of human subjects, human interactions, balanced reporting, and fiscal responsibility. Taken as a group, the propriety standards require that evaluations be conducted legally, ethically, and with due regard for the welfare of those involved in the process and those affected by the results.

The fourth group, which includes the *accuracy* standards, determines whether an evaluation has produced sound information. These standards require that the obtained information be technically adequate and that conclusions and recommendations be logically linked to the obtained information. The eleven concerns treated in this area are object identification, context analysis, described purposes and procedures, defensible information sources, valid measurement, reliable measurement, systematic data control, analysis of quantitative information, justified conclusions, and objective reporting. The evaluation's overall rating against the accuracy standards is intended to reflect its overall validity.

In order to address each of the thirty concerns in detail, each standard is elaborated to explain why it is important, to suggest how it might best be met, and to illustrate the consequences of violating it in a given situation. While the standards contain operational advice, they are also flexible and allow for creativity in evaluations. Moreover, they are written in plain English and have been judged to be useful and of high quality in both hearings and field tests. The Joint Committee has recommended that the ***Standards*** be consulted in relation to a wide range of tasks in evaluation. These include deciding whether to do an evaluation, contracting, design,

staffing, collection of information, analysis, reporting, research, and training. Hence, the standards have a practical, as well as a philosophical, orientation.

To assure their long-term usability, the ***Standards*** must be subjected to an ongoing program of review and revision. Accordingly, the Joint Committee plans to become a standing committee with evolving membership. It has also arranged for the royalties from the sales of the ***Standards*** to be deposited into a trust fund that will be used to support ongoing review, revision, and associated developmental activities. This plan is consistent with those that guide the setting of standards in more mature professions such as accounting, auditing, and educational testing.

In addition to the Joint Committee's efforts, it will be important for other persons involved in evaluating education to help in the review and revision effort. I hope very much that major journals in all areas of education that are concerned with the conduct and use of evaluation will soon carry reviews of the ***Standards.*** Along these lines, Dr. Barbara Davis, editor of the ***Evaluation News,*** has planned a series of articles designed to assess the ***Standards*** from various perspectives. In addition to conceptually-oriented reviews, it will be important to obtain feedback about the strengths and weaknesses of the ***Standards*** from persons who use them. The Joint Committee, therefore, plans to invite feedback and to study and use it in the process of revising the ***Standards.***

It should be kept in mind that the main point of professional standards is not merely to have them but to implement them. The educational evaluation profession as a whole, which includes its members, sponsors, and clients, should use the standards to promote evaluation practice that serves the cause of education and the pertinent needs of society. Such groups can encourage the sound use of the standards by engaging in training, research, and development related to the use of the standards. These are constructive ways of achieving widespread implementation and, as a general policy, are to be preferred to the negative approach that emphasizes enforcement and sanctions. Educational evaluators should recognize that if they do not effectively regulate their activities, other groups, especially courts and governmental bodies, will. In general, those concerned with advancing the practice of educational evaluation should carefully consider how they can promote and ensure the implementation of the standards.

Research

In the abstract, the need for research on educational evaluation need not be explained and argued at length. Most will agree that the vitality and growth of any field is highly dependent on sustained efforts to produce new knowledge about relationships among crucial variables in the field. In particular, persons closely involved with educational evaluation would likely agree that we know far too little about what goes on under the label of educational evaluation and about which practices work and which ones do not. Moreover, I doubt that anyone would argue that educational

evaluators are as actively involved as they should be in researching their field.

The ***Standards*** present an opportunity for educational evaluators to promote and guide research on their field. As a whole, the ***Standards*** may be viewed as a framework that defines important variables in educational evaluation, as well as possible relationships among them. In particular, the individual guidelines may be considered working hypotheses about what constitutes good evaluation and what methods are powerful for achieving desirable evaluation outcomes. The ***Standards*** book should prove to be a valuable resource for persons who wish to develop and implement an agenda of research on educational evaluation.

In reviewing the content of the ***Standards,*** I have identified the six areas for research. They are described below.

State of the Art Studies

The first area concerns the state of the art in educational evaluation. Now that the field has articulated 30 standards of good practice, it will be important to ascertain the quality of evaluations in relation to each of them. Such information should help to identify which standards should be given priority attention in efforts to advance the practice of evaluation. In addition, studies that search for problems and promising practices, irrespective of the topics covered by the 30 standards, are needed. These will help to ensure that the full range of important issues are investigated. They will also help those persons concerned with reviewing and revising the 30 standards to search for important areas that may not be sufficiently addressed at present.

A few questions that might be addressed in state of the art studies are listed below.

1. Given information about compliance with the requirements of the 30 standards for specified samples of completed or ongoing evaluations, how good are the evaluations in relation to each standard?
2. Given information about the compliance with the requirements of the 30 standards, for specified samples of different types of evaluations (e.g., small vs. large, formative vs. summative, internal vs. external, and program vs. project vs. materials), which types rate best on which standards?
3. Do research designs patterned after different evaluation models vary systematically in their adherence to the requirements of the 30 standards?
4. What are the relative merits of alternative techniques (such as oral and printed reports) for meeting certain standards; for example, evaluation impact, clear reporting, timely reporting, and objective reporting?
5. What would practicing evaluators and evaluation clients identify as the most important strengths and weaknesses in the practice of educational evaluation?

The preceding questions are only illustrative of the needed state of the art studies.

Validity of the Evaluation Standards

A second area of needed research concerns the validity of the claims that are implicit and explicit in the ***Standards***. There are many such claims in the ***Standards***. Some of them promote certain evaluation practices while others advise against applying the ***Standards*** to certain kinds of evaluations. Both types of claims have great bearing on the validity of the advice offered in the ***Standards***. To ensure that the ***Standards*** do not provide faulty advice and that their full potential is realized, their underlying claims must be identified and tested. In accordance with the findings, the content of the ***Standards*** and the advice offered for their use should then be modified.

The nature of the claims to be tested is illustrated by the following questions.

1. Are the 30 standards equally applicable to evaluations of educational programs, projects, and materials?
2. Are the standards unbiased in regard to the premises and requirements of particular evaluation models?
3. Are the standards equally applicable to evaluations that are internal as well as external, small as well as large, and formative as well as summative?
4. Are the standards applicable to educational endeavors generally or only to those in the United States?
5. Are the standards applicable only to evaluations of programs, projects, and materials, or might they also be sufficient to guide and assess evaluations of such objects as personnel and institutions?
6. Are the standards applicable only to evaluations in education or might they also be sufficient to guide and assess evaluations in such areas as social work, medicine, energy, and ecology?
7. Are some standards consistently more important than others in determining the quality of evaluations?
8. In general, does adherence to the guidelines, pitfalls, and caveats presented with each standard increase the likelihood of meeting it?

The above questions reflect some of the claims in the standards that should be tested in order to assess and strengthen the validity of the advice they contain.

The Process Underlying the Standards

A third area of needed research concerns the process of applying the ***Standards***. Thus far there is only limited experience in implementing them. The field tests that have been conducted indicate that people who apply the ***Standards*** in different contexts and in relation to different kinds of evaluations will encounter a wide range of issues.

Some of the questions about application, which have been identified thus far, include the following.

1. What criteria and procedures are useful in determining the relative importance of different standards in given evaluations?
2. What guidance can be offered for resolving conflicts between the standards?

3. What pre-evaluational analysis may be useful to reach a decision about whether the standards will be met sufficiently to warrant proceeding with an evaluation?
4. What criteria and information are appropriate for assessing the extent to which given standards have been met?

In order to clarify the ***Standards*** and develop helpful recommendations for their implementation, it will be important to investigate questions such as those listed above. In general, such investigations should be designed to identify and assess recurrent problems and the ways that evaluators have responded to them.

Uses and Abuses of the Standards

A fourth area of needed research involves the uses and abuses of the ***Standards***. The authors of the ***Standards*** have suggested a wide range of potential uses that could result in a general upgrading of evaluation practice. In addition, reactors to the ***Standards*** have speculated that they might be misused in a variety of ways. These projections signal the need for studies to examine how the ***Standards*** are actually used and with what effect.

Listed below are questions that bear on this topic.

1. To what extent and with what effect are the standards used to guide the following evaluation tasks:
   a. Deciding whether to evaluate
   b. Designing evaluations
   c. Contracting for evaluations
   d. Staffing evaluations
   e. Managing evaluations
   f. Collecting information
   g. Analyzing information
   h. Reporting results
   i. Applying results
   j. Evaluating evaluation studies
   k. Certifying evaluators
   l. Suing for malpractice in evaluation
   m. Researching the evaluation process
2. Which standards are considered most extensively in performing the tasks listed above?
3. What are the arguments for and against applying the standards to the tasks listed above?

Research on questions about the uses of the ***Standards*** is essential for developing and issuing recommendations that promote appropriate uses and warn against misuses.

Costs and Benefits of Applying the Standards

Closely aligned with investigations of uses and abuses of the ***Standards*** are questions concerning the costs and benefits of applying them. Several questions are pertinent to this area.

1. How extensively are the standards used?
2. On the average, what added costs are incurred when applying the standards?
3. How does use of the standards to plan and conduct evaluations affect the quality of evaluations and the audiences' use of evaluation results?
4. How does use of the standards to plan and conduct evaluations affect the evaluations' credibility?
5. What are the unanticipated effects of applying the standards?
6. Do independent groups who must plan similar studies arrive at a more uniform evaluation approach when they apply the standards than when they do not?
7. Do the standards facilitate agreement among judges about the quality of given evaluations?
8. What are the comparative effects of applying the Joint Committee's standards and other related sets, such as those developed by the Evaluation Research Society of America?

Efforts to systematically apply the standards, to monitor their use, and periodically to update them will obviously involve a significant investment of time and resources. In order to justify this investment, it will be essential to identify and weigh the actual costs and benefits as illustrated by the questions listed above.

Issues Highlighted in Given Standards

The final area of suggested research pertains to particular questions that may be derived from given standards. For example, the "valuational interpretation" standard identifies five ways of assigning value meanings to findings. According to that standard, interpretations may be based on: (1) the extent to which objectives are achieved, (2) an analysis of judgments gathered from many people about the meaning of the findings, (3) the evaluator's judgment, (4) the relation of the findings to assessed needs, or (5) the client's judgment. Whereas the "valuational interpretation" standard claims no superiority for any of these approaches, different leaders in the educational evaluation field have urged the use of one or another of these approaches above the others Research on this point—and on others that pertain to given standards— would be useful for validating or modifying the advice offered in the individual standards.

The main points of this section are listed below.

1. Further development of the evaluation profession requires an intensified effort to conduct research on the evaluation process.
2. The recently released ***Standards*** provide a useful conceptual framework for guiding such research.
3. Six general areas of research seem to be indicated.
4. Study of the six areas raises a number of questions illustrating the nature of particular investigations that evaluation researchers might soon pursue.

In concluding this section, I recommend that evaluation researchers increase their efforts to develop scientific understanding of the evaluation

process and that they seriously consider the role of the ***Standards*** in guiding the needed program of research.

Training

Earlier in this paper, I argued that training in educational evaluation is vital to ensure the professional stature of the field. I further charged that evaluation training in education is grossly inadequate. In this section I will elaborate on these claims, propose some ways of looking at the training problem, and offer a few suggestions.

Training in educational evaluation is important for several reasons. First, the involvement of a wide range of persons is necessary for sound application of evaluation. Second, these persons need the competence and confidence to perform their roles. Third, these roles often require specialized skills. And, fourth, educational evaluators need to keep pace with the rapidly expanding state of their profession. Since the formalized practice of evaluation is crucial in education, a sound and powerful program of pertinent training is essential.

Sadly, the present training program is weak. In fact, it hardly can be defended as a program, in other words, as a systematic and goal-directed process. Conceptualizations of educational evaluation training are almost non-existent; the audiences for training are not clearly identified; skill needs are discussed only vaguely in the evaluation literature; and, to my knowledge, there are no carefully articulated strategies for the delivery of training to the variety of persons who participate in the evaluation of education.

I recognize that there currently are some important training activities in educational evaluation. These include graduate courses and degree programs at a few universities; on-the-job training opportunities in some labs, centers, government agencies, and school districts; pre-sessions and post-sessions at some conventions; a variety of other workshops and institutes; a network of Title I technical assistance centers; and some self-instructional materials. These are important resources on which to build, but they do not add up to the needed program of training.

A key reason why the present training efforts are inadequate, I believe, is the limited effort by federal and state agencies to fund curriculum development and training in educational evaluation. The Bureau of Education for the Handicapped and the Title I programs are major exceptions to this charge. Overall, though, state education departments, NIE, NSF, and the Office of Education have done little to promote and support a sound program of evaluation training. This seems odd and misguided since these agencies invest heavily in educational evaluation. I hope that government agencies become convinced that launching and sustaining an extensive program of evaluation training is in their interest and in the public interest.

Another reason for the inadequacy of evaluation training is that the audience for such training typically has been too narrowly defined. Most training is focused on evaluation specialists. The many other persons, whose participation is needed to ensure the success of evaluation, remain

largely untrained and unsocialized in the process of evaluation. Thus, many evaluation efforts have failed because the audiences didn't know how to assess and use its results.

If I had to choose the audience for my evaluation training efforts, I would not hesitate to focus my attention on future educational administrators. If they come to value, understand, and not fear the process of evaluation, they are likely to promote its use, support it with funds, focus it on their agency's problems, and use its results. Otherwise, they are likely to resist evaluation, and the efforts of evaluation specialists in their districts will likely make little difference.

It is not necessary, however, to focus evaluation training exclusively on any one group. Instead, evaluation trainers can and should focus their training on all persons who participate in conducting evaluations and applying their results. In general, this audience may be divided into three levels: (1) evaluation specialists, including especially the directors and staff members in evaluation agencies; (2) evaluation generalists, including especially superintendents, principals, teachers, and counselors; and (3) lay groups concerned with education, including school boards, advisory councils, PTA's, and parents. In rough terms, the number of persons who need evaluation training at each of these levels is inversely proportional to the amount of training they need. For example, while there are a great many lay persons who need to be trained, their training needs are small compared to those of the lesser number of administrators and teachers; and, in turn, their requirements are modest compared to the extensive training needs of the relatively small group of evaluation specialists. Yet, all of these groups need to learn a common language about evaluation, and they need to learn how to work together toward common ends.

As I think about how these diverse groups might be trained to work together in evaluations, I cannot help but return to some hunches I followed in my early days as an evaluation trainer. One was that some, but not all, training should be done in the context of real evaluations. Another was that training should be offered to all participants in an evaluation. One more was that all participants should be given a common set of terms and definitions that they can use to communicate about evaluation. And still another was that the main thrust should be toward helping all trainees to conceptualize the evaluation process and its application to a variety of problems. In the main, these hunches point toward field-based training, but they also call for grounding in theory.

I strongly recommend field-based training as a strategy. One reason is that it forces a consideration of relevance, as well as rigor, in an evaluation. Another is that it confronts the trainees with the full range of problems in evaluations. Moreover, it helps a variety of persons to consider their evaluation roles vis-à-vis those of other participants. Finally, every evaluation presents a potential laboratory setting for evaluation training. For these reasons, I think it is a good idea, whenever feasible, to incorporate training activities into the design of an evaluation. If this were a common practice, I suspect tha. the dissemination of evaluation

expertise throughout educational agencies would dramatically increase.

And because evaluation is largely a conceptual and communication process, I also see the need for grounding in evaluation theory. All participants in an evaluation need such grounding, at least at a general level. In other words, evaluation trainers should offer a kind of core curriculum in evaluation theory to all of the groups that participate.

Until recently, this need was nearly intractable because of the great diversity in the evaluation literature, the jargon-laden presentations, and the focus on the needs of evaluation specialists. This situation should change dramatically with the publication of the ***Standards*** for several reasons. First, they represent, I believe, widely-shared agreements about educational evaluation; second, they are addressed to all participants in such evaluations; third, they are plainly written; fourth, they have a practical, as well as a theoretical orientation; and, fifth, there will be an ongoing effort to review and improve them. While the ***Standards*** won't meet all of the content needs in evaluation training, I believe they provide a powerful base on which to build.

This concludes my discussion of the need for increased and improved training in educational evaluation. I hope I've convinced you that this need should be given a high priority as evaluators plan to improve their profession.

Conclusion

In this paper, I've tried to respond to Wayne Welch's charge by explaining why I believe that educational evaluation is a vital field. I've acknowledged that this field is primitive and in need of improvement. I've pointed to the Evaluation ***Standards*** as a resource that can be used to guide some of the needed improvement efforts. And I have argued that research and training in educational evaluation are crucial and in need of vast improvement. If nothing else, I hope I have helped you to think about the future of evaluation and to rededicate your efforts to improve it.

About the Contributors

W. James Popham is a professor in the Graduate School of Education at the University of California, Los Angeles. He is also Director of the International Objectives Exchange, a Los Angeles-based agency that develops educational assessment devices. Popham is a past President of the American Educational Research Association and edits ***Educational Evaluation and Policy Analysis***. He has received numerous federal grants dealing primarily with the assessment of instructional competency in the classroom. In addition to writing several books, he has authored many research papers, articles, and monographs in the areas of testing and educational evaluation.

Daniel L. Stufflebeam is a professor on the Faculty of Educational Leadership and the Director of The Evaluation Center at Western Michigan University. During more than fifteen years of educational research experience, Stufflebeam has developed numerous standardized achievement tests, several evaluation systems, and the CIPP Evaluation Model. He has published extensively in the area of educational evaluation and has served on the editorial panels of such publications as the ***Journal of Higher Education***, ***Evaluation Comment***, and Evaluation Center Occasional Paper series. During the past four years, he has headed the development of the Standards for Evaluation of Educational Programs.

Nick L. Smith is the Director of the Research on Evaluation Program at the Northwest Regional Educational Laboratory. He served as President of the Evaluation Network in 1980. Smith has written numerous articles on educational evaluation, many dealing with the development of new approaches and methods. He is editor of the Research on Evaluation Program quarterly newsletter and has served as a reviewer for the ***Journal of Educational Measurement***, ***American Educational Research Journal***, and ***Educational Evaluation and Policy Analysis***.

Michael Scriven is the Director of the Evaluation Institute at the University of San Francisco and edits the newsletter ***Evaluation Notes***. He is a past President of the American Educational Research Association and of the Evaluation Network. Scriven is a prolific author in philosophy, science, technology, as well as educational evaluation. He is also the author of ***Evaluation Thesaurus*** and of two evaluation books soon to be published.